Strengthening Collaborative Data Analysis and Decision Making in Web Communities

Nikos Karacapilidis
Computer Technology Institute &
Press "Diophantus" and
University of Patras
26504 Rio Patras, Greece
nikos@mech.upatras.gr

Spyros Christodoulou
Computer Technology Institute &
Press "Diophantus" and
University of Patras
26504 Rio Patras, Greece
shristod@cti.gr

Manolis Tzagarakis
Computer Technology Institute &
Press "Diophantus" and
University of Patras
26504 Rio Patras, Greece
tzagara@upatras.gr

Georgia Tsiliki
Bioinformatics and
Medical Informatics Team
Biomedical Research Foundation
Academy of Athens, Greece
gtsiliki@bioacademy.gr

Costas Pappis
Dept. of Industrial Management
and Technology
University of Piraeus
18534 Piraeus, Greece
pappis@unipi.gr

ABSTRACT

Generally speaking, modern research becomes increasingly interdisciplinary and collaborative in nature. Researchers need to collaborate and make decisions by meaningfully assembling, mining and analyzing available large-scale volumes of complex multi-faceted data residing in different sources. At the same time, they need to efficiently and effectively exploit services available over the Web. Arguing that dealing with data-intensive and cognitively complex settings is not a technical problem alone, this paper presents a novel collaboration support platform for Web communities. The proposed solution adopts a hybrid approach that builds on the synergy between machine and human intelligence to facilitate the underlying sense-making and decision making processes. User experience shows that the platform enables stakeholders to make better, more informed and quicker decisions. The functionalities of the proposed platform are described through a real-world case from a biomedical research community.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *data sharing, web-based services;* H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – *collaborative computing, computer-supported cooperative work, web-based interaction.*

General Terms

Management, Design, Experimentation, Human Factors.

Keywords

Computer-supported cooperative work on the Web, collaborative decision making, data mining, Web intelligence, Web communities, Integration, Web services.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2578845

1. INTRODUCTION

Collection and analysis of large quantities of data are foundational challenges in all fields of science and engineering [1]. At the same time, research in the majority of these fields has become increasingly interdisciplinary and collaborative in nature [2]. In such settings, where each individual member of a Web community can easily contribute to the data tsunami, data may vary in terms of subjectivity and importance, ranging from individual opinions and estimations to broadly accepted practices and well-documented scientific results. In addition, data types can be of diverse level as far as human understanding and machine interpretation are concerned.

The above remarks advocate the exploitation of the synergy between human and machine reasoning when designing systems to support such processes within Web communities. Exploitation of data mining technologies for pattern and dependencies discovery within large data sets is certainly of great benefit. However, in spite of big progress made in the area of computational analysis, there are many patterns that humans can easily detect but computer algorithms struggle to estimate [1]. Interpretation of analysis' results is a challenging issue here, in that getting results from the execution of a data mining algorithm is rarely enough; additional information is needed concerning how each result came out and based on which input.

This paper presents an innovative Web-based collaboration support platform, which has been developed to address the above issues. Arguing that dealing with data-intensive and cognitively-complex settings is not a technical problem alone, the proposed solution has been developed in the context of an FP7 EU research project, namely Dicode (http://dicode-project.eu), and fully embeds data mining in a collaborative data analysis and decision making context. Specific functionalities of the Dicode platform are described through a real-world case concerning a biomedical research community. Such communities have to deal with large-scale amounts of multiple types of data, obtained from diverse and distributed sources, while a vast growth of publicly available biomedical resources are available on the Web. In addition, recent technology advances, such as those in Next Generation Sequencing (NGS) platforms, entail an exponential increase in the size and number of experimental data sets available [3].

2. RELATED WORK

Generally speaking, the emergence of the Web 2.0 era introduced a plethora of collaboration tools which enable massive scale engagement and feature novel paradigms. These tools cover a broad spectrum of needs that range from file sharing to mind-mapping and argumentative collaboration support. Yet, such tools are generic and - in most cases - very difficult to interoperate; thus, while useful in supporting team work, their separate use becomes prohibitively expensive.

Focusing on the biomedical research domain, a series of applications and Web services that link together bioinformatic tools and databases have recently emerged, showing the way to easily analyze biomedical data. For instance, BioGRID (http://thebiogrid.org) is a repository that stores readily combined data sets and provides platforms to easily visualize such data; the GenePattern platform provides access to more than 180 tools for genomic analysis to enable reproducible in silico research (http://www.broadinstitute.org/cancer/software/genepattern). Integration of these separate systems and resources into a single flexible infrastructure that streamlines heterogeneous workloads is still a challenging task.

At the same time, a number of biomedical research related projects and initiatives aim at addressing diverse collaboration requirements in a variety of contexts. For instance, GRANATUM (http://granatum.org) tries to bridge the information, knowledge and collaboration gap by providing integrated access to the globally available data resources needed to perform complex cancer chemoprevention experiments and conduct studies on large-scale datasets; SIMBioMS (http://simbioms.org) is a multi-module solution for biomedical data management that is able to accommodate experiments requiring non-conventional data storage solutions. While certainly helpful in addressing specific biomedical subjects, the above projects and initiatives do not deal with ‘big data’ issues; also, they do not exploit the synergy between human and machine intelligence in order to meaningfully accommodate and interpret the results of the associated data mining services through an environment that facilitates and enhances collaboration among stakeholders.

As the number of related Web services is constantly increasing, their proper integration becomes a critical issue. A few approaches have been already launched to facilitate the collaboration, data sharing and decision making among scientists by providing them with a platform to share resources. A well known example of this category of related work is myExperiment (http://www.myexperiment.org), an online research environment that supports the social sharing of bioinformatics workflows, i.e. procedures consisting of a series of computational tasks, which can then be shared and reused according to their specific requirements. Another representative example is BioCatalogue (http://www.biocatalogue.org), which is a registry of Web services that allows users to annotate and comment on the available services in order to assist them in identifying the more suitable ones. In any case, approaches of this category demonstrate a set of limitations, mainly concerning incorporation of collective intelligence and flexibility in the integration of services offered. Moreover, they lack mechanisms for a meaningful integration of data mining services to appropriately support tasks such as the discovery of patterns and dependencies within large data sets, which are very common in the biomedical research domain.

3. A BIOMEDICAL RESEARCH CONTEXT

The context under consideration concerns multidisciplinary biomedical research communities, with members ranging from biologists to bioinformaticians, which need to collaborate in order to assimilate clinico-genomic research information and scientific findings and explore diverse associated issues. Recent studies indicate that a culture where every scientist needs to understand how to manage, navigate and curate large-scale data needs to be developed [4].

In this context, biomedical researchers often augment their in-house data with publicly available data stored in varying formats. A typical process is to download the raw or the pre-processed data from a database (e.g. Gene Expression Omnibus) along with all the relevant phenotypical and clinical information needed to understand and analyze the data. The analysis could be conducted by using either a standalone tool, such as Cytoscape (www.cytoscape.org), or in-house scripting using, for example, the R statistical language. In any case, the most important step in the life cycle of an experiment is to interpret and communicate the findings. Specifically, findings need to be meaningfully interpreted to have an insight into the initial biological question of interest. For that purpose, researchers confer with databases, such as the Kyoto Encyclopaedia of Genes and Genomes, or standalone tools which are directly linked to databases and can qualitatively and quantitatively assess the submitted results using the database resources. Decision making plays an important role here; scientists need to evaluate their options of analyses, databases, tools, and often base their decisions on past experience and feedback from their colleagues.

In fact, the above context concerns scientific data exploration and analysis, in which scientists may - individually or collaboratively - exploit heterogeneous data sources, available tools, algorithms or services. A thorough analysis of a specific biomedical research assimilator context, performed in the early stages of the Dicode project and involving 8 senior researchers with diverse background, revealed the need for development of innovative solutions that enhance interdisciplinary collaboration and decision-making by facilitating information integration under a common platform. Such a solution should enable stakeholders to identify data and annotation databases, share their own experiences and findings, externalize their tacit knowledge, efficiently handle large amounts of data, share predictive models for data analysis, exploit a set of data mining algorithms that are tailored to biomedical research needs, share and collaboratively interpret the outcomes of the data mining algorithms, monitor data and decision provenance issues, and build a social network for effective interaction.

In the specific biomedical research assimilator context reported in this paper, data sources incorporated concern four different types of data, as shown in Table 1. The table concerns data related to the breast-cancer disease, but it could be easily generalized to other diseases or organisms. To give an indication of the data scale associated to the context under consideration, representative numbers of samples and data sizes are also given.

As results from the above, a holistic approach integrating collaboration, decision making and data mining services is required. The approach described in the next section is geared towards this direction.

Table 1. Input data considered for the biomedical research assimilator context

Data type & description	Databases (Web available)	Data in numbers
Genomics/ Transcriptomics data: Normalized or raw data	Gene Expression Omnibus	86 datasets; 7,607 samples (~ 500Kb per sample, ~ 32Mb per dataset)
	ArrayExpress	978 experiments; 69,483 samples
	Stanford Microarray Database	508 experiments
Phenotypic data: Supplementary, clinical or phenotypic data available	As above	2 files on average per dataset (~10Kb per dataset)
Molecular Pathways: Data from known and established molecular networks	Kyoto Encyclopedia of Genes and Genomes	416 pathway maps (153,758 total)
	Reactome	3,931,211 data entries
Annotation data: Reference databases for biomedical & genomic information	Gene Ontology	~ 30,000 terms; ~ 50,000 relationships
	National Center of Biotechnology Information	26,473 annotated coding regions (RefSeq); 129,493 homo sapiens entries (UniGene); ~127 billion bases (GenBank); > 21 million citations for biomedical literature (PubMed)

4. THE DICODE APPROACH

The overall goal of the Dicode project is to facilitate and augment collaboration and decision making in diverse data-intensive and cognitively-complex settings. To do so, whenever appropriate, it builds on prominent high-performance computing paradigms and large scale data processing technologies to meaningfully search, analyze and aggregate data existing in diverse, extremely large, and rapidly evolving sources. At the same time, particular emphasis is given to the proper exploitation and analysis of large scale data, as well as to collaboration and sense making support issues. Building on current advancements, the solution offered by the Dicode project brings together the reasoning capabilities of both the machine and the humans. It enables the meaningful incorporation and orchestration of a set of interoperable Web services that reduce the data-intensiveness and complexity overload of the settings under consideration to a manageable level, thus permitting stakeholders to be more productive and effective in their work practices.

Services that have been developed and integrated for the context under consideration include: (i) *data acquisition services* that enable the capturing of tractable information existing in diverse data sources and formats, (ii) *data mining services* that provide functionality such as looking for subgroups in any user-provided data and recommending similar users or documents from log data, (iii) *collaboration support services* that facilitate the synchronous and asynchronous collaboration of stakeholders through adaptive workspaces, efficiently handle the representation and visualization of the outcomes of the data mining services, and enable the orchestration of a series of actions for the appropriate handling of data, and (iv) *decision making support services* that exploit a series of reasoning mechanisms to enhance both individual and collective sense- and decision-making.

Central to the proposed approach is the concept of the Dicode Workbench, which refers to a Web-based application that enables the seamless integration of heterogeneous services, including advanced data mining and collaboration services, while it ensures the interoperability of these services from both a technical and a conceptual point of view. In this regard, semantics techniques have been exploited to define an ontological framework for capturing and representing the diverse stakeholder and services perspectives. Figure 1 illustrates an instance of the Dicode Workbench. As shown, a widget-like approach has been adopted, where each widget implements a particular Web service. The Workbench can be personalized, in the sense that an end-user may add or remove widgets (e.g. according to the needs of the particular context and issue under exploration). The central widget of Figure 1 hosts the collaboration and decision making support service (which will be further analyzed in the next section), while widgets on the right and left side host various data acquisition and data mining services. The Dicode Workbench allows users to maximize any of the widgets located on the sides; when a widget is maximized, it changes its position with the widget that is in the middle at that moment (thus reflecting where the focus of attention each time is).

The Dicode Workbench enables integration of services in two distinct types, either at the user interface level, or at a deeper, semantic level. In the former, the integrated services are displayed and simply coexist on the same Web page but no exchange of data or any other interaction takes place between them. To achieve this type of integration, services need only to provide a REST-based interface, which is invoked by the Workbench to trigger their execution. In the second integration type, the integrated services do not only coexist on the same Web page, but are also able to exchange data for a particular purpose (this is described in detail in the next section). This integration type supports user-friendly functionalities, such as 'drag-and-drop' for passing data from one service to another. In this case, the Dicode Workbench actually constitutes the communication channel to enable data exchange and interaction between services. To support this integration type, apart from what is required in the case of the previous one, a loosely coupled architecture has been implemented based on the idea of passing message interfaces (MPI) [5].

5. COLLABORATION IN DICODE

Being fully integrated into the Dicode Workbench, collaboration and decision making support services enable participants to collectively reflect on various issues, their ultimate aim being to jointly decide about which course of action to take.

5.1 Conceptual Approach

Collaboration in Dicode brings together two paradigms: the Web 2.0 paradigm, which builds on flexible rules favoring ease-of-use and human interpretable semantics, and the traditional decision support paradigm, which requires rigid rules that reduce ease-of-use but render machine interpretable semantics. To achieve this, our approach builds on a conceptual framework, where formality and the level of knowledge structuring during collaboration is not considered as a predefined and rigid property, but rather as an adaptable aspect that can be modified to meet the needs of the tasks at hand. Allowing formality to vary within the collaboration space, *incremental formalization*, i.e. a stepwise and controlled evolution from a mere collection of individual ideas and resources

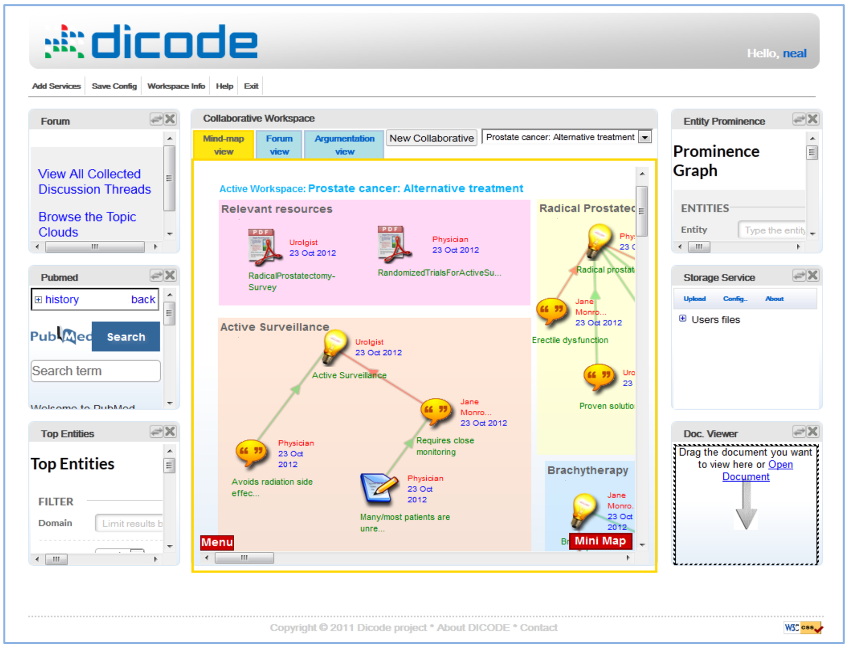

Figure 1: An instance of the Dicode Workbench.

to the production of highly contextualized and interrelated knowledge artifacts and finally decisions, can be achieved.

Dicode offers alternative visualizations of the collaboration space (called 'Dicode views'), which comply with the incremental formalization concept. Each Dicode view provides the necessary mechanisms to support a particular level of formality. The more informal a view is, the greater easiness-of-use is implied. At the same time, the actions that users may perform are intuitive and not time consuming; however, the overall context is human (and not system) interpretable. On the other hand, the more formal a view is, the smaller easiness-of-use is rendered; the actions permitted are less and less intuitive and more time consuming. The overall context in this case is both human and system interpretable [6].

The functionality described in this paper is offered through the Dicode 'mind-map view', in which a collaboration space is displayed as a mind map (Figure 2), where users can upload and interrelate diverse types of items (a detailed description of these items is given in Section 5.2). This view deploys a spatial metaphor permitting the easy movement and arrangement of items on the collaboration space. Stakeholders may organize their collaboration through dedicated item types such as *ideas*, *notes*, *comments* and *services*. Ideas stand for items that deserve further exploitation; they may correspond to an alternative solution to the issue under consideration and they usually trigger the evolution of the collaboration. Notes are generally considered as items expressing one's knowledge about the overall issue, an already asserted idea or note. Comments are items that usually express less strong statements and are uploaded to express some explanatory text or point to some potentially useful information. Finally, service items enable users to configure, trigger and monitor the execution of external services from within the workspace, and allow the automatic upload of their results into the workspace (as soon as the execution of the service is completed). Multimedia resources can also be uploaded into the 'mind-map view'.

5.2 Scenario of Use

To better illustrate the use of the proposed Web-based collaboration support platform, this subsection presents an illustrative scenario concerning collaboration in the area of breast cancer research (a recording of the platform's use appears at http://dicodedev.cti.gr/www2014/paper.html).

Alice is a Pharmacology Ph.D. student. Her research is on adjuvant hormonal therapy for patients with breast cancer disease; particularly, she is interested in identifying how Tamoxifen (Tam) resistant cells modulate global gene expression. Tam is a widely used antagonist of the estrogen receptor (ER), whereas its resistance is a well-known obstacle to successful breast cancer treatment [7]. Alice selected and analyzed gene-expression data from 300 patient samples with the help of Neal, an MD at a collaborating university hospital, and Jim, a postdoctoral researcher in Bioinformatics. These data are derived from whole human genome expression arrays (Affy U133A Plus 2.0 – see http://www.affymetrix.com). Although the sample is relatively large, Alice believes that augmenting the data with publicly available data will be a good idea for statistically significant results.

To analyze the data and discuss the analysis results, Alice, Neal and Jim decide to collaborate by using the Dicode mind-map view. In this direction, Alice is launching a new collaborative workspace (Figure 2). Even though all three collaborators are aware of the benefits and difficulties of Tam treatment, Alice adds a note on the collaboration workspace to fully explain the characteristics of the genomic data (Figure 2, (a)). Neal has collected all the necessary clinical information and posts them on the collaboration space (Figure 2, (b)). Apart from stating the scientific question of interest, Alice summarizes the biological background and technical difficulties around it (Figure 2, (c)), while Neal finds an interesting article concerning the Tam treatment and uploads the corresponding pdf file on the workspace (Figure 2, (d)). In the mind-map view, users may

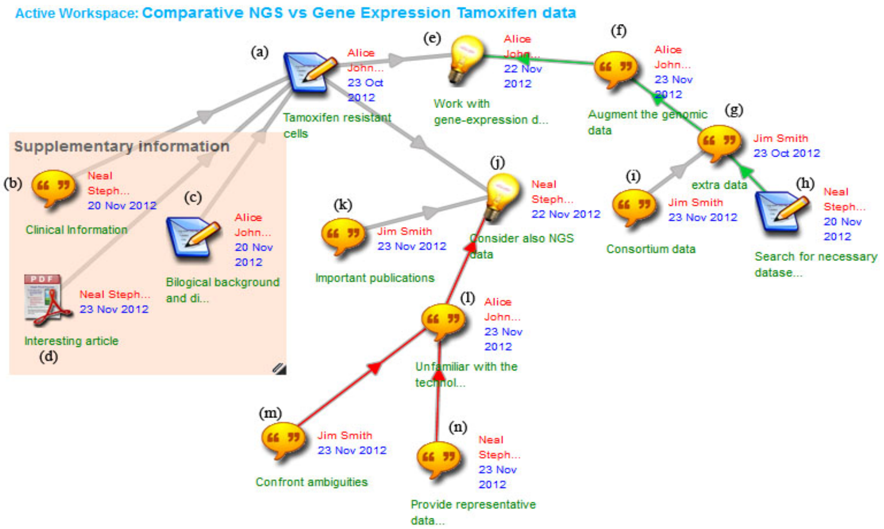
Active Workspace: Comparative NGS vs Gene Expression Tamoxifen data
(a)
Alice John... 23 Oct 2012
Tamoxifen resistant cells
(e)
Alice John... 22 Nov 2012
Work with gene-expression d...
(f)
Alice John... 23 Nov 2012
Augment the genomic data
(g)
Jim Smith 23 Oct 2012
extra data
(h)
Neal Steph... 20 Nov 2012
Search for necessary datase...
(i)
Jim Smith 23 Nov 2012
Consortium data
Supplementary information
(b)
Neal Steph... 20 Nov 2012
Clinical Information
(c)
Alice John... 20 Nov 2012
Bilogical background and di...
PDF
Neal Steph... 23 Nov 2012
Interesting article
(d)
(j)
Neal Steph... 22 Nov 2012
Consider also NGS data
(k)
Jim Smith 23 Nov 2012
Important publications
(l)
Alice John... 23 Nov 2012
Unfamiliar with the technol...
(m)
Jim Smith 23 Nov 2012
Confront ambiguities
(n)
Neal Steph... 23 Nov 2012
Provide representative data...

Figure 2: Launching a collaboration workspace for estimating the dominance of Tamoxifen resistant cells to global gene expression.

group together related items by using colored rectangles (see, for instance, the one entitled "Supplementary information").

Alice believes that they should first work with the gene-expression data (idea item (e), Figure 2) and, moreover, they should augment the genomic data (comment item (f), Figure 2). Jim suggests launching the GEO_Recommender (GEOR) service (Figure 2, (g)) to find "similar" data sets in terms of pathology characteristics. GEOR is a Web service implemented in the context of Dicode, which searches the GEO database (http://www.ncbi.nlm.nih.gov/geo) based on keywords or the description supplied by the user. Having that in mind, Neal offers to find the extra data sets (Figure 2, (h)), since he is more confident with the technical characteristics of the data. Jim agrees (Figure 2, (i)), and adds that there are data available from consortiums such as caBIG (https://cabig.nci.nih.gov), which have extensively proven the need to augment or at least compare and assess findings across multiple data sets.

Even though Alice believes that they should first work with the gene-expression data, Neal argues that they should also consider NGS data (idea item (j), Figure 2). He mentions that he is responsible for a clinical trial and can have access to total RNA from human breast cancer cell lines, which are then analyzed using NGS technology. Jim is also working with NGS data and he is highly recommending the integration or at least the comparative study of the two platforms (Figure 2, (k)). Moreover, NGS is the latest technology having higher specificity and sensitivity, and thus has higher potential in meaningfully augmenting Alice's results.

Alice is reluctant to start working with NGS data because she is unfamiliar with the technology and argues that she will probably invest time without being assured about the significance of the results (Figure 2, (l) – note that arrows in red denote argumentation against the 'father' item, while arrows in green denote argumentation in favor). To defeat this statement, Neal suggests (Figure 2, (m)) to provide her with a representative data set from his laboratory, while Jim offers to help her (Figure 2, (n)) deal with all the ambiguities between the two datasets.

Alice thinks about exploiting the *Subgroup Discovery* (SD) data mining algorithm [8], which is offered through a Dicode service. SD is the task of finding patterns that describe subsets of a data set that are highly correlated relative to a target attribute. This is a popular approach for identifying interesting patterns in the data, since it combines a sound statistical methodology with an understandable representation of patterns. Through a dedicated service interface, Alice proceeds in entering the required parameters that include the input file containing the data, the number of rules to be used, the service ontology, as well as the list of attributes to be included/excluded. Once these parameters are entered, Alice starts the execution of the SD service (this invokes the REST-based API of the SD service).

Upon the successful termination of the service's execution (Figure 3(a)), the service outcomes are automatically uploaded on the collaboration workspace (Figure 3, (c)). One collaboration item is created for each result returned (in this scenario, outcomes are in html format). Actually, these results are tables with GO (http://www.geneontology.org) and KEGG (http://www.genome.jp/kegg) terms, which describe biological processes related to the estimated groups of genes. For this particular run, the SD returns four subgroups (Results 1-4, Figure 3, (c)). The results of the SD service seem convincing to Neal (Figure 3, (d-e)), while Jim expresses his opposition about the third outcome and quotes a part of a scientific paper he recently read (Figure 3, (f-g)).

The same procedure (invoking the SD service and collectively assessing its output) is then followed for the NGS data (Figure 3, (h) and (j)). The three researchers carefully examine the commonalities between the two SD runs (on genomic and NGS data) and share their insights. The subgroups returned for the NGS data ((Figure 3, (i)) are very similar to the ones obtained from SD service on genomic data. Alice is impressed with the commonalities found between the two SD runs; she is now convinced that there is scope to integrate additional NGS data. She expresses her insight (Figure 3, (k)) and links it to the original Neal's idea (note that SD service items are also linked as arguments in favor of this insight). To further elaborate this issue,

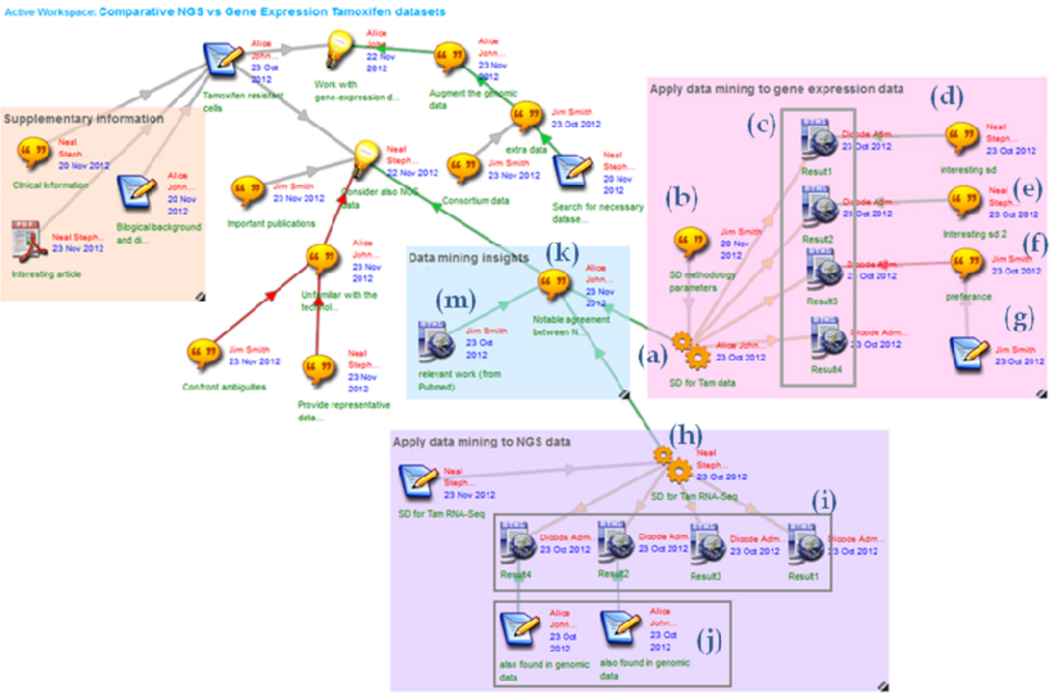
Active Workspace: Comparative NGS vs Gene Expression Tamoxifen datasets
Alice John... 23 Oct 2012
Tamoxifen resistant cells
Alice John... 22 Nov 2012
Work with gene-expression d...
Alice John... 23 Nov 2012
Augment the genomic data
Jim Smith 23 Oct 2012
extra data
Supplementary information
Neal Steph... 20 Nov 2012
Clinical Information
Alice John... 20 Nov 2012
Biogical background and di...
Neal Steph... 23 Nov 2012
Interesting article
Neal Steph... 22 Nov 2012
Consider also NGS data
Jim Smith 23 Nov 2012
Important publications
Jim Smith 23 Nov 2012
Consortium data
Neal Steph... 20 Nov 2012
Search for necessary datase...
Alice John... 23 Nov 2012
Unfamiliar with the technol...
Jim Smith 23 Nov 2012
Confront ambiguites
Neal Steph... 23 Nov 2012
Provide representative data...
Data mining insights
(k)
Alice John... 23 Nov 2012
Notable agreement between N...
(m)
Jim Smith 23 Oct 2012
relevant work (from Pubmed)
Apply data mining to gene expression data
(a)
Alice John... 23 Oct 2012
SD for Tam data
(b)
Jim Smith 20 Nov 2012
SD methodology parameters
(c)
Dicode Adm... 23 Oct 2012
Result1
Result2
Result3
Result4
(d)
Neal Steph... 23 Oct 2012
interesting sd
(e)
Interesting sd 2
(f)
Jim Smith 23 Oct 2012
preferance
(g)
Apply data mining to NGS data
(h)
Neal Steph... 23 Oct 2012
SD for Tam RNA-Seq
Neal Steph... 23 Nov 2012
SD for Tam RNA-Seq
(i)
Dicode Adm... 23 Oct 2012
Result4
Result2
Result3
Result1
(j)
Alice John... 23 Oct 2012
also found in genomic data

Figure 3: Application of SD service to NGS data, assessment of results, and insights.

Jim uses the PubMed service offered through the Dicode Workbench to search for recent relevant articles. He then uploads a link (Figure 3, (m)) pointing to a scientific report that strengthens Alice's argument. The above collaboration may proceed to further augment the gene expression and NGS data.

6. CONCLUSIONS

The Dicode platform is able to fully exploit the synergy between human and machine reasoning. In cases where biomedical knowledge discovery has to be based on data mining tools, the platform enables stakeholders to set up a collaborative, interactive process, where they can easily decide about which data repositories should be considered, trigger and parameterize the associated data mining mechanisms, explore their discovery patterns, discuss the weaknesses of the identified patterns, control the complexity of the output, and set up new iterations of the data mining algorithm by defining other attributes or considering alternative data. It is also noted that the Dicode platform fosters standards-based integration and exploitation of information resources and data analysis tools across organizational boundaries. Dicode has been developed by adopting existing standards and custom Web technology. The platform can be easily customized, through a proper assembly of Web services and associated data resources.

As a concluding note, we argue that the Dicode approach enables a meaningful aggregation and analysis of large-scale data in complex settings, such as that of biomedical research. The proposed solution allows for new working practices that turn the problem of information overload and cognitive complexity into the benefit of knowledge discovery. It thus improves the quality of collaboration within a Web community, while enabling its users to be more productive and focus on creative activities.

7. REFERENCES

[1] Challenges and Opportunities with Big Data. White Paper, Computing Community Consortium. Spring 2012, http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

[2] Lee, E. S. 2007. Facilitating collaborative biomedical research. In GROUP '07 Doctoral Consortium papers, pages 5:1–5:2, New York, NY, USA, 2007. ACM.

[3] Brazas, M.D, Yamada, J. T. T. and Ouellette, B. F. 2009. Evolution in bioinformatic resources: 2009 update on the Bioinformatics Links Directory. Nucleic acids research, 37:W3–5.

[4] Haendel, M.A., Vasilevsky, N.A., Wirz, J.A. 2012. Dealing with Data: A Case Study on Information and Data Management Literacy. PLoS Biol. 10, 5, e1001339.

[5] Snir, M., Otto, S. W., Walker, D. W., Dongarra, J., and Huss-Lederman, S. 1995. MPI: The Complete Reference. MIT Press, Cambridge, MA, USA.

[6] Karacapilidis, N. and Tzagarakis, M. 2012. Towards a Seamless Integration of Human and Machine Reasoning in Data-Intensive Collaborative Decision Making Settings: The Dicode Approach. In Proc. of the 16th Int. Conf. on Decision Support Systems, IOS Press, Amsterdam, 223-228.

[7] Huber-Keener, K.J., Liu, X., Wang, Z. et al. 2012. Differential Gene Expression in Tamoxifen-Resistant Breast Cancer Cells Revealed by a New Analytical Model of RNA-Seq Data. PLoS ONE. 7, 7, e41333. DOI= doi:10.1371/journal.pone.0041333.

[8] Atzmueller, M., Puppe, F. and Buscher, H.P. 2005. Exploiting background knowledge for knowledge-intensive subgroup discovery. In Proc. of IJCAI'05, 647-652.